Econ 270C: Analytics of Economic Development
Problem Set 1 (Due Tuesday April 4, 2006)

Child Health and Education: Non-parametric regressions

For this problem, download the STATA dataset "PSET1.DTA" from the course page (http://emlab.berkeley.edu/users/webfac/emiguel/e270c_s06/e270c.shtml). The data are from a joint project with Michael Kremer and Rebecca Thornton in rural Kenya (on the course syllabus, "Incentives to Learn", although note that the data you will use here is only a subset of the overall dataset). Please include all regression output and graphs, as well as do-files, with your solutions.

This problem examines the relationship between merit awards and academic performance, as measured by school exams, among Kenyan schoolchildren. In early 2001, Grade 6 girls in a random subset of "treatment" schools (variable name "treat") were offered a large cash award if they scored in the top 15% of all treatment school girls. The dataset contains test score information from late 2000, the year before the program, and late 2002, one year after the program had ended. The test score outcomes ("test00", "test02") were normalized such that the test distribution in the comparison schools is mean zero with a standard deviation of one (for all students in that grade, not just those in this sample – thus the mean need not equal 0).

The goal of this problem set is to understand medium-term impacts of the program using various parametric and non-parametric methods. Another important issue for the analysis is whether girls at the bottom of the baseline test score distribution were harmed by the program – perhaps due to demoralization or diversion of teacher attention to high-achieving classmates, for instance.

a) Present summary statistics for the three variables in the dataset. Do baseline 2000 test scores differ on average across the treatment and comparison students? Do 2002 test scores differ on average across treatment groups? **[1 point]**
(STATA command hints: "summarize, detail", "ttest")

b) Plot the kernel density of 2000 test scores in the following ways:

(i) Epanechnikov kernel with the optimal bandwidth
(ii) Epanechnikov kernel with bandwidth equal to 0.05
(iii) Gaussian kernel with optimal bandwidth

Characterize the distribution of 2000 test scores. Does kernel density estimation appear to be more sensitive to the bandwidth or to the kernel? **[1 point]**
(STATA command hints: "kdensity", "graph twoway")

c) Plot the kernel density of 2000 test scores with Epanechnikov kernel and optimal bandwidth separately for treatment and comparison students, and place them in one figure.

Do the same for 2002 scores. Describe any shifts in the distributions through time. **[1 point]**

d) Determine the linear relationship between the 2002 test score (dependent variable) and the 2000 test score using OLS, and present the regression results.

Plot the predicted quadratic relationship between the 2002 test score and the 2000 test scores, as well as 95 percent confidence bounds.

Plot the predicted quadratic relationship between 2002 and 2000 test scores separately for treatment and comparison students, and place them in one figure. **[1 points]**
(STATA command hints: "regress", "graph", "qfit / qfitci")

e) Perform the following non-parametric Lowess regressions of 2002 test score (dependent variable) on 2000 test score:

(i) All students, bandwidth equal to 0.5
(ii) All students, bandwidth equal to 0.05
(iii) Treatment students, bandwidth equal to 0.5
(iv) Comparison students, bandwidth equal to 0.5

Characterize the non-parametric relationship between the 2002 test score and 2000 test score. Does the relationship appear to differ for treatment versus comparison students? Are there important patterns that were not apparent from the linear or quadratic fits in part (d)? **[1 point]** (STATA command hints: "lowess", "graph, twoway")

f) Perform the Fan locally-weighted non-parametric regression, using a quartic kernel with bandwidth equal to 0.5. Include 95 percent confidence bounds in all plots, bootstrapping the standard errors.[1] "Trim" extreme values from the distribution, only considering the 2000 test score interval of [-1.3, 1.6], for simplicity.

(i) Treatment pupils, regress the 2002 test (dependent variable) on the 2000 test.
(ii) Comparison pupils, regress the 2002 test (dependent variable) on the 2000 test.
(iii) Take the difference between these two non-parametric relationships, and bootstrap the standard errors of this difference, for various points throughout the 2000 test score distribution.

Is the 2002 difference between treatment and comparison students significantly different than zero (at over 95 percent confidence) anywhere in the 2000 test score distribution? **[2 points]**

g) Estimate the reduced-form impact of treatment assignment on 2002 test scores using OLS (taking into account any common error component within schools). Is this relationship robust to controlling for the baseline (2000) test score? Do any subgroups of girls, by initial age ("age"), gain more than others from the program? Present the regression results and discuss. **[1 point]** (STATA command hints: "cluster")

h) What are the implications of these results for our understanding of the relationship between merit awards, child effort, and learning in rural Kenya? Are most test score gains at the "top" of the distribution, in the "middle", or at the "bottom"? Is there any conclusive evidence of negative externalities for low-achieving students?

Perform any additional econometric analysis (with the current dataset) that you feel may shed light additional on this issue of heterogeneous program impacts. **[2 points]**

---

[1] Refer to *The Analysis of Household Surveys* by Angus Deaton for the STATA code on Fan regressions and bootstrapping. These are also online at *www.worldbank.org/LSMS/tools/deaton*. Be aware that there are a few errors in the published Fan regression code that you will need to spot.